下载APP

19 | 协同过滤：你看到的短视频都是集体智慧的结晶

2021-09-10 郭炜

《数据分析思维课》 课程介绍 >

讲述：郭炜
时长 16:43 大小 15.32M

数据给你一双看透本质的眼睛，这里是《数据分析思维课》，我是郭炜。

学到这里，其实你已经了解了不少复合算法的使用，包括和时间相关的马尔可夫链，加速我们进行选择的蒙特卡洛和拉斯维加斯算法等等。

在实际算法的应用过程当中，还有一种通过集体智慧来构成的复合算法，它可以寻找大量人群当中的行为数据模型规律，达成普通算法从单体上无法达到的效果。这种算法当中，最著名的一个算法就是协同过滤算法。

协同过滤算法定义与场景

协同过滤算法顾名思义，就是指用户可以齐心协力，通过不断地和算法互动，在多如牛毛的选择当中，过滤掉自己不感兴趣的选择，保留自己感兴趣的选择。

协同过滤算法源于 1992 年，最早被施乐公司发明并用于个性化推送的邮件系统（施乐公司就是那个发明了 GUI 界面，被乔布斯发现并创造了 MAC OS 的公司）。最早这个算法是让用户从几十种主题里面去选 3~5 种自己感兴趣的主题，然后通过协同过滤算法，施乐就根据不同的主题来筛选人群发送邮件，最终达到个性化邮件的目的。

到 1994 年的时候，协同过滤算法开始引入集体智慧的概念，也就是用更多的人群和数据去获取相关的知识。它允许用户贡献自己的一些行为和反馈，从而创造一个比任何个人和组织更强大机制，自动给用户发送喜欢的文章。

基于这个思路，施乐发明了著名的 GroupLens 系统。在这个系统里面，用户每读完一条新闻都会给一个评分，系统会根据这些评分来确定这些新闻还可以推送给谁。你看，今日头条的想法其实施乐在 1994 年就实现了。

当然当时的算法还没有那么精准，不过这个系统完全颠覆了过去只能通过编辑人工规则推送文章的机制，它是基于用户自己的每次反馈自动找到和该用户类似的人，再发送新闻邮件。通过 GroupLens 系统大获成功之后，协同过滤算法迅速占领了推荐系统的市场。因为**推荐系统需要同时具备速度快和准确度高两个特点（需要在用户打开网站几秒钟就要推荐所感兴趣的内容或者物品），而协同过滤算法正好满足了这两点要求，这也是这个算法经久不衰的原因。**

使用协同过滤算法里最著名的网站就是亚马逊的网络书店，你每次去选择一本你喜欢感兴趣的书籍，马上就会看到下面有关于“浏览此商品的顾客也同时浏览”的推荐。

又比如我们去逛 B 站，你会看到 B 站会根据你自己的资料和类似的人浏览的视频来帮你找到可能感兴趣的视频，例如你喜欢二次元，你看到的推荐大部分都是二次元的视频；像我这样技术宅的，给我推荐的就都是各种技术类型的视频。

这样协同过滤系统从单一系统内的邮件文件过滤到跨系统的新闻电影，再到最后我们看到的短视频。虽然它们推荐的东西不同，但给我们的体验都是类似的：你总是能看到你喜欢

的产品、感兴趣的服务、喜欢的视频、想读的文章。

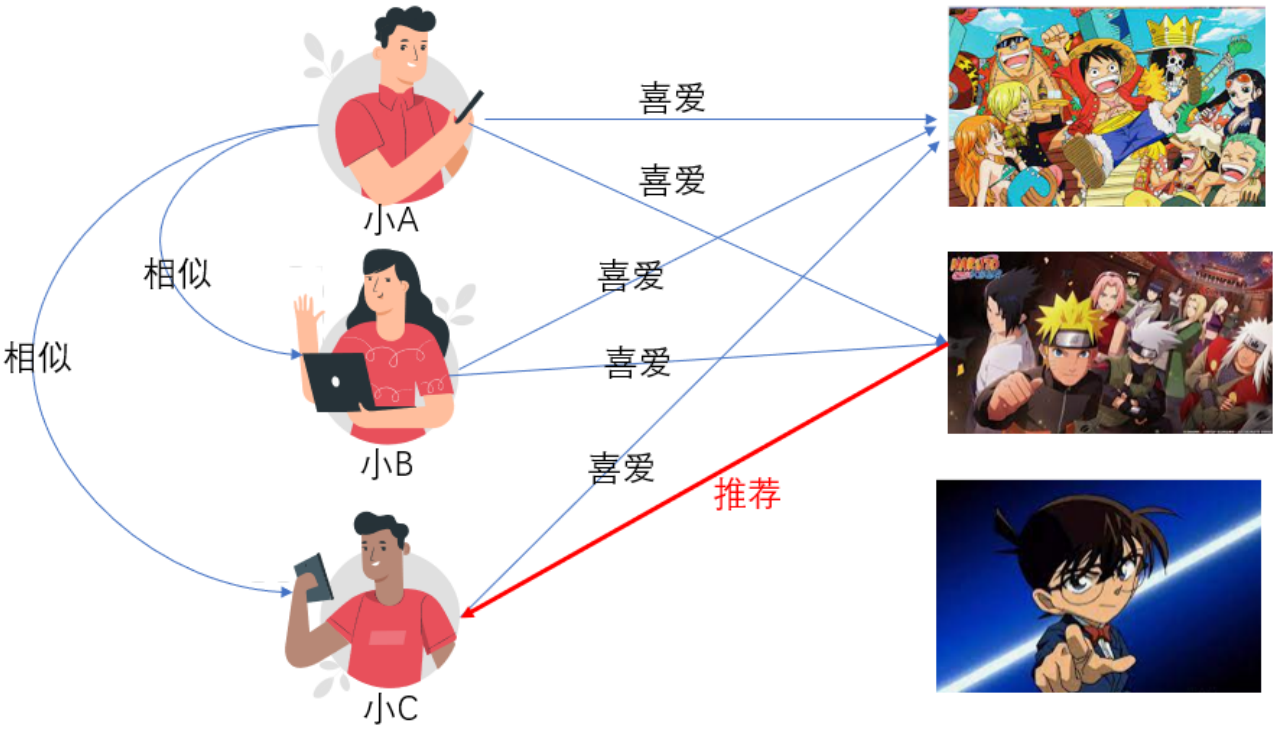
协同过滤算法简述

协同过滤算法是怎么样能够知道你所喜欢的东西呢？我来给你介绍三个最常见的协同算法，它们分别是：**基于用户的协同过滤算法**（User-based Collaborative Filtering）、**基于物品的协同过滤算法**（Item-based Collaborative Filtering）以及**基于数据模型的协同过滤算法**（Model-based Collaborative Filtering）。

基于用户的协同过滤算法

基于用户的协同过滤算法就是基于用户和用户之间的相似性，推荐你喜欢的东西，过滤你不喜欢的东西。

例如你喜欢看海贼王，然后和你相似度很高的一个用户他喜欢火影忍者。我们应该把火影忍者推荐给你吗？还不能确定。因为只有一个和你类似的人喜欢火影忍者，不代表着你就会喜欢火影忍者。协同过滤算法是有和你相似的 N 个人都喜欢火影忍者，算法就会把火影忍者推荐给你们这群人。就像下面这个图，首先得先找到和你有类似喜好的用户，然后根据你喜欢和这些这类用户都喜欢的物品来给你推荐。



在下面这个场景里小 A 小 B 小 C 都喜欢海贼王，根据用户协同过滤算法公式判断小 A 小 B 小 C 相似距离小于阈值；而小 A 和小 B 都喜欢火影忍者，因此根据基于用户的协同过滤的算法，我们把火影忍者也推荐给小 C，认为他也喜欢看火影忍者。

基于用户的协同过滤在实际的使用当中为了提高生产效率，还会用到一些倒排的算法、数据特征选取（例如选取冷门商品而非热门商品，比如都要买高考习题集锦并不一定代表着用户相关性，但是喜欢看宅舞的用户相似度应该很高）、还有特征权重算法等等。

基于用户的协同过滤在使用当中有优点也有缺点，它的优点在于：

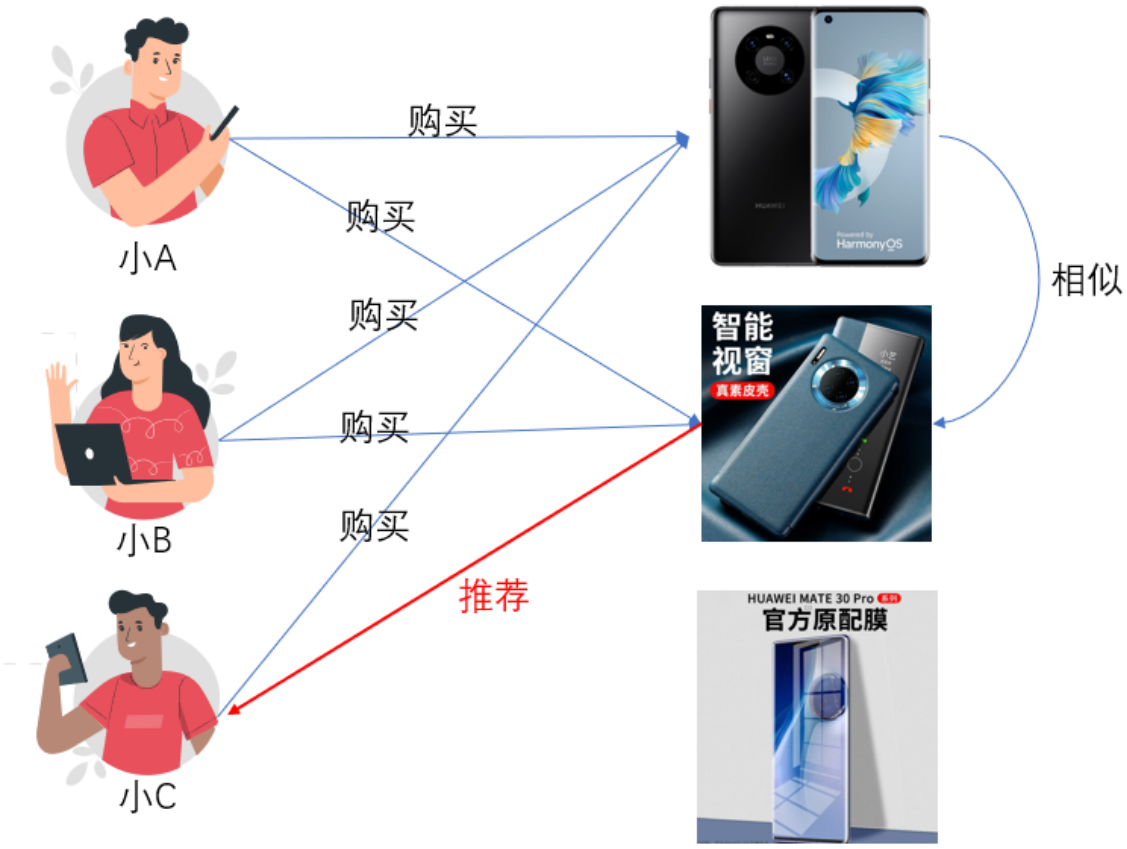
- 基于用户的协同过滤是找到用户之间的相似程度，所以能够反映一些小群体当中的物品的热门程度；
- 它可以让用户发现一些惊喜。因为是根据类似用户的喜好来推荐，用户会发现自己对一些过去不知道的东西是感兴趣的；
- 对一些新的有意思的物品比较友好。一旦某一个新的商品和电影被某一个社群的用户购买了，我们马上就可以推荐给他圈子当中的其他用户。

它的缺点在于：

- 如果你是一个新的用户，你可能不能马上找到和你类似的人，所以无法马上获得准确的推荐；
- 对于推荐出来的结果虽然会给你带来惊喜，但是它也不太有解释性；系统不知道推荐给你的这个物品是什么，只知道你的相关的朋友都在使用；
- 对于用户群比较大的公司，去计算用户之间的相似度的话，计算的耗费会比较高。

基于物品的协同过滤算法

接下来我们来看看基于物品的协同过滤算法。**这种算法就是根据用户群对于物品的购买或者评价发现物品和物品之间的相似程度，然后再根据具体用户历史使用的类似物品推荐给这个用户。**这么说可能有些绕，我举个例子你就明白了。比如华为的手机和华为手机壳经常被一起购买，这两个物品之间就存在比较强的相关性。那么当一个用户去新购买一个华为手机的时候，我们就会给他推荐一个华为的手机壳。整体的推荐逻辑和算法如下面这个图。



在这个场景里小 A 小 B 在购买华为手机的时候都购买了华为手机壳，而根据物品协同过滤算法公式判断华为手机和华为手机壳的相似距离小于阈值。那么小 C 在买华为手机的时候，根据基于物品的协同过滤的算法，我们把华为手机壳也推荐给小 C，认为他也可能购买华为手机壳。

基于物品的协同过滤算法同样也要计算物品之间的相似程度，也要用到一些倒排算法和物品的特征选取以及过滤噪音的算法，它的优点是：

推荐更加针对用户自身。它反映了每个用户自己的兴趣的决策，根据你自己每买的一个商品来给你做推荐，而不是一类人给你做推荐；

实时性比较高。用户每次点赞和购买商品都可以对其他购买此商品的用户推荐；

推荐的结果很好解释。因为它们都是类似或者是关联度很高的商品，推荐结果显而易见。

这种算法的缺点在于：

对于新加入进来的商品反馈速度比较慢，因为没有人购买也没有人互动，所以可能有一些很好的商品没有被很好地推荐；反过来没被推荐买的人更少，推荐的可能性更低，出

现“产品死角”；

不会给用户惊喜：大部分的物品其实都是关联度比较高，可以被想象到，惊喜程度不高；

对于商品或物品更新情况比较快的领域比较不适用，比如新闻。因为你没有推荐到别人看，可能这个新闻就过期了。

具体在生产环节使用的时候，一般对小型的推荐系统来讲，基于物品的协同过滤是大家使用的主流。因为基于物品的协同过滤算法计算量比较小，上手程度比较快。随着你的用户变得更多、用户的期待更高，一般都会过渡到基于用户的协同过滤算法。

基于数据模型的协同过滤算法

我们很容易能想到，不一定非要通过协同过滤的公式来计算用户之间的距离，**我们完全可以复用前面所学到的算法，先做出来模型，再进行相关的协同过滤。**

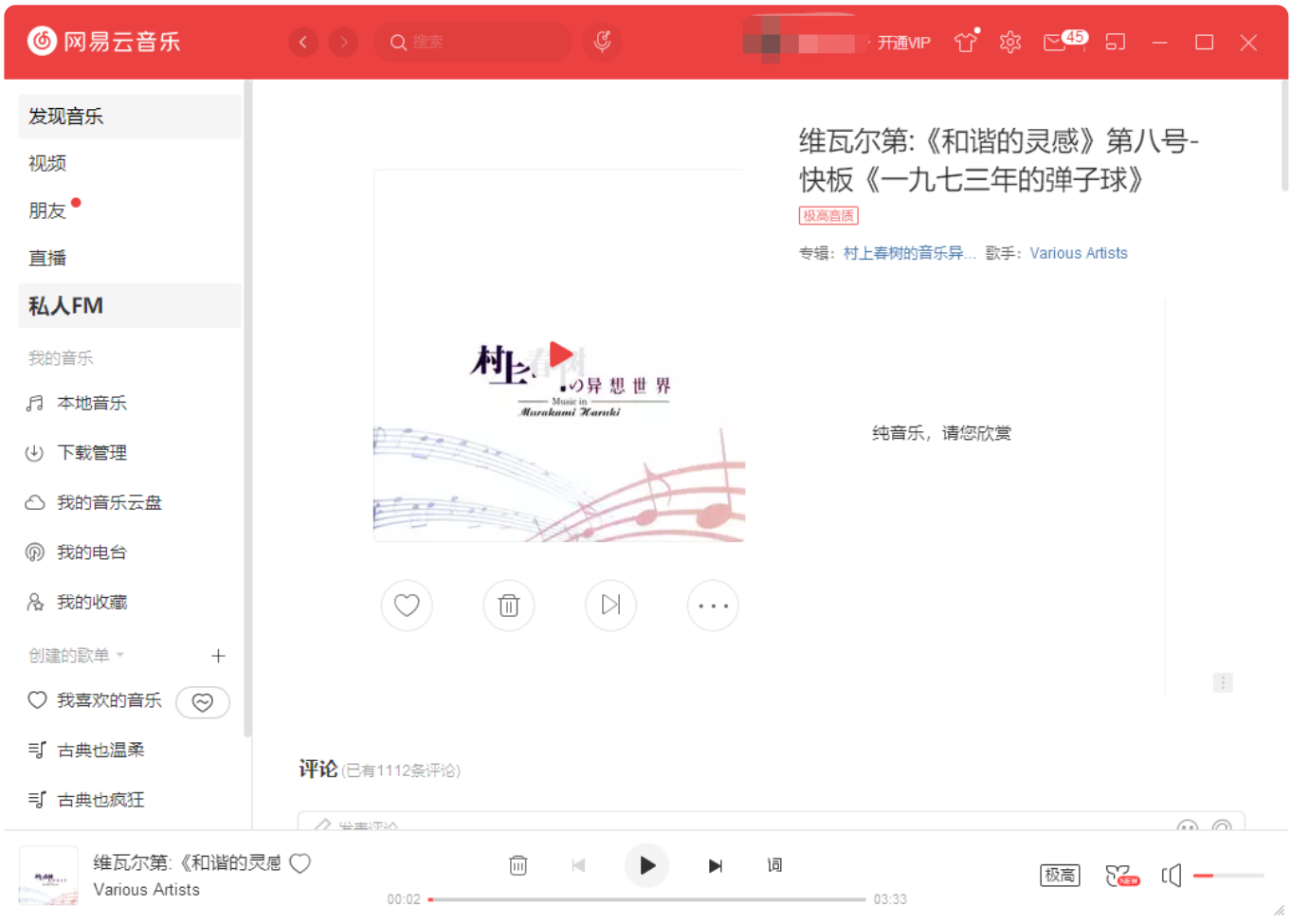
例如，用关联算法来去做物品之间的相似度评估，然后根据置信度、支持度、提升度或者其他评分规则推荐给用户。我们也可以用聚类算法来找到用户之间的关联程度，把用户之间的距离计算出来，然后把这些用户群相似度比较高的用户之间的商品确定给同样聚类比较清楚的用户。类似的我们还可以用分类算法回归算法，用神经网络做协同过滤、用图模型做协同过滤、用隐语义模型做协同过滤等等。

所以协同过滤它是一个利用“集体智慧”复合算法的思想，可以用前面我们介绍的所有类似的算法找到物品和物品或者用户和用户之间的关系来去做协同过滤。

协同过滤的使用场景与缺点

协同过滤算法已经被应用到互联网的方方面面，通过这些基于协同过滤算法的推荐算法，在 2016 年的双十一，阿里巴巴平台取得了 20% 的增长；Youtube 上 70% 的用户时长都是协同过滤算法贡献的；Netflix 75% 的播放也来自推荐系统，帮助 Netflix 每年节省 10 亿美元的广告费，可见协同过滤算法多么的强大。

同样，你在网易音乐听到的私人 FM 这也是通过协同过滤的推荐算法来实现的，它可以给你推荐你从来不认识的歌手，但你听这个歌手的音乐也会很喜欢。



不过协同过滤算法并不是万能的。因为它是考虑集体智慧给你带来新鲜的推荐和最适合感兴趣的东西，因此对具体某一个个体的实际情况并没有更多的判断。

例如你会经常看到，你浏览手机之后购买了一个手机，但是接下来几周，某些电商网站里还是会坚持不懈地给你推荐手机，这是因为协同过滤不是针对你个体进行的推荐。同时，因为基于集体智慧，所以对于一些很冷门的商品来说，最初始的这些用户流量是需要有引导的，否则会导致有些很不错的商品一直放在角落里无人问津。这对一些优秀但冷门的产品，例如小制片人的电影其实就不是特别友好。

同样因为协同过滤只是考虑到了物品和用户之间的关系，没有考虑到用户所处的场景，所以推荐的内容可能就不太有效。例如在上班的时间给你去推荐一些餐厅，但你那个时候并不能去；你带着孩子出去游玩，开车的路上不停给你推荐一些课程，这其实都只是考虑到了用户之间物品之间的关系，没有考虑到场景这个点。这就需要数据分析师、算法科学家结合具体的业务场景和实际所有的数据再进行特殊的算法优化，这样协同过滤在场景当中才能有更高的使用性。

而使用协同过滤的算法最大的弊病在于，**这个算法就像是一个溺爱你的妈妈，永远会给你想要的东西，它并没有价值观，你会被“惯”得越来越没有节制，把时间全都花费到各种各样的短视频、小文章、和你钟爱的小圈子里，但最终你并没有什么拓展和收获。**它不是一个严肃的爸爸，能告诉你你应该去学什么、哪些价值观是对的，这些其实是协同过滤算法无法做到的，只有通过人的选择和经验才能告诉你，你应该去学什么，而不是拼命满足你自己的某些爱好。

对于这方面的缺陷，现在有不少科学家尝试通过深度学习的方法模拟价值观、人类的思考来进行修正，这个算法会不断地精进。

小结

今天给你整体讲了基于集体智慧的协同过滤算法。它最大的价值就是打破了过去的一个我们常见的规则，就是前面讲过的帕雷托定律（也就是二八法则）：通常是 20% 的大品牌占据了 80% 的市场，而小品牌正占据剩下的 20%。

协同过滤根据每一个人自己的品牌偏好充分传播和扩展，让长尾品牌“聚沙成塔”。协同过滤让和主流有所不同的一些小众的品牌，慢慢地让喜好的人群去接触到。抖音、头条等把一些很小众但很有特色的视频和新闻带给最需要的人，这打破了我们过去看到的以主流

流量为主导的新闻体系，将长尾效应发挥到了极致，改变了原来市场上的规则。这也是为什么抖音和头条在新浪、网易、搜狐等巨头垄断的情况下还可以发展壮大的原因。

协同过滤算法也给我们很多启示。

首先，你自己的心态应该更加开放，不要一股脑地追主流，毕竟主流和大众的不一定是最适合自己的，我们的圈子当中应该有个性化的东西。

同样，我们的价值观也应该更加地开放，不能就沉浸在自己的小圈子里。因为协同过滤给我们的都是我们所喜欢的东西，它的价值观并不一定是最好的，我们应该开放心态去接受和尝试各种各样新的主流的非主流的物品，用我们自己的经历和人生去判断。

我们更不要沉浸在某些短视频或者网站根据我们兴趣推荐的碎片化文章里。因为它给我们带来的不是推荐，而是去束缚、固化我们的思维，让我们成为这个时代里的“井底之蛙”。毕竟我们要主导自己的人生，而不是让算法去主导我们的人生。

数据给你一双看透本质的眼睛。协同过滤把集体智慧带入了算法世界，同时，我们也要警惕，因为推荐只是推荐，要持续去扩宽你的思维边界。

课后思考

协同过滤是一个“人人为我，我为人人”的集体智慧算法，在你的生活当中，你有观察到哪些事情和协同过滤算法比较类似吗？分享出来，我们一起提高。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

赞 3 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 18 | 马尔可夫链：你的未来，只取决于你当下做什么

下一篇 20 | 人工智能初探：阿尔法狗是怎样的一只“狗”？

精选留言 (3)

写留言

那时刻 神

2021-09-10

技术改变我们的生活，带来更多的便利。但是不能主导我们的生活。我们应该把多一点碎片化的时间用来学习，学习这些技术的本质。

展开

1

范

2021-09-14

被“溺爱”也是一种奴役。系统通过各种协同过滤算法宠爱人类，人类也还是要保留自己的思考。

进化菌 神

2021-09-11

协同过滤算法，就好像我们选择团队比较合适的代码框架，往往选的是大多数人熟悉的，而不是新颖独特的。
不过，协同的弊病是导致过分同类化，让人更容易缺乏尝试新鲜事物，扩展性确实不太好。进而，热门的事物持续热门，冷的造致爆冷，这就有点像有些优秀的作品不被大多数人看见~

展开